On the Leakage of Personally Identifiable Information Via Online Social Networks

Balachander Krishnamurthy, AT&T Labs – Research
Craig E. Wills, Worcester Polytechnic Institute

ACM SIGCOMM Workshop on Online Social Networks

Barcelona, Spain

August 2009

Personally Identifiable Information

PII defined as information which can be used to distinguish or trace an individual's identity either alone or when combined with other public information that is linkable to a specific individual.

Users provide various pieces of PII to Online Social Networks (OSNs), which is often visible to more than just friends.

Third-party servers for aggregating user viewing behavior are prevalent in popular Web sites as well as OSNs.

Key Question: is PII belonging to any user being leaked to these third-party servers via OSNs?

Overview of Results

PII leakage does occur.

Happens because of PII leakage via HTTP headers sent to third-party aggregators.

Most users on OSNs are vulnerable to having their OSN identity information linked with tracking cookies.

Have shared this information to all the OSNs we studied so that they may make informed decisions regarding preventative measures and subscriber notification.

The goal of this work is not a legal examination of privacy policies, but to bring a technical examination of the observed leakage to the community's attention and to propose means to prevent such leakage.

Consequences

1. With tracking cookies having been set and gathered for several years to track user visits to *non-OSN* sites as well, it is possible for third-party aggregators to associate identity with those *past* accesses.
2. As users on OSNs will continue to visit OSN *and* non-OSN sites, such behavior in the *future* is also liable to be linked with their OSN identity.

Aggregators claim they create profiles of users based on their Internet behavior, but do not gather or record PII.

Although we do not know that aggregators are recording PII, we demonstrate with this work that it is undeniable that information *is* available to them—either directly or indirectly via OSN identifiers.

Availability of PII in OSNs

Examined a total of 12 OSNs for the pieces of PII that are requested of users (note users do not necessarily supply all of this information or do so truthfully).

Bebo, Digg, Facebook, Friendster, Hi5, Imeem, LinkedIn, LiveJournal, MySpace, Orkut, Twitter and Xanga.

Availability of PII Pieces to Non-Friends in 12 OSNs

	Level of Availability			
Piece of PII	Always Available	Available by default	Unavailable by default	Always Unavailable
Personal Photo	9	2	1	0
Location	5	7	0	0
Gender	4	6	0	2
Name	5	6	1	0
Friends	1	10	1	0
Activities	2	8	0	2
Photo Set	0	9	0	3
Age/Birth Year	2	5	4	1
Schools	0	8	1	3
Employer	0	6	1	5
Birthday	0	4	7	1
Zip Code	0	0	10	2
Email Address	0	0	12	0
Phone Number	0	0	6	6
Street Address	0	0	4	8

Count of OSNs shown; rows go from bad to good wrt privacy concerns.

Most users use default permissive settings.

Leakage Detection Methodology

Used *Live HTTP Headers* extension for Firefox browser to capture complete HTTP header information while interacting with each of 12 OSNs studied.

In each case examined if and how OSN identifier is leaked to third-party aggregators.

Sample leakage (via an embedded object on `myspace.com` page):

```
GET /pagead/test_domain.js HTTP/1.1
Host: googleads.g.doubleclick.net
Referer: http://profile.myspace.com/index.cfm?
     fuseaction=user.viewprofile&friendid=123456789
Cookie:id=2015bdfb9ec||t=1234359834|et=730|cs=7aepmsks
```

Four Types of PII Leakage

1. transmission of the OSN identifier to third-party servers from the OSN;
2. transmission of the OSN identifier to third-party servers via popular external applications;
3. transmission of specific pieces of PII to third-party servers; and
4. linking of PII leakage within, across, and beyond OSNs.

Leakage occurs via Referer Header, Request-URI and Cookie.

Leakage of OSN Identifier

```
GET /clk;203330889;26770264;z;u=ds&sv1=170988623...
Host: ad.doubleclick.net
Referer: http://www.facebook.com/profile.php?
     id=123456789|&ref=name
Cookie: id=2015bdfb9ec||t=1234359834|et=730|cs=7aepmsks
```

Leakage of Facebook identifier to doubleclick.net via Referer header.

Observed OSN id being leaked via OSN via at least one header for 11 out of 12 OSNs studied (and Orkut, the 12th, is operated by Google).

Leakage Via External Applications

```
GET /track/?...&fb_sig_time=1236041837.3573&
     fb_sig_user=123456789&...
Host: adtracker.socialmedia.com
Referer: http://apps.facebook.com/kick_ass/...
Cookie: fbuserid=123456789;...=blog.socialmedia.com...
     cookname=anon; cookid=594...074; bbuserid=...;
```

Leakage of Facebook identifier to socialmedia.com via Request-URI and Cookie.

Leakage of Pieces of PII

```
GET /st?ad_type=iframe&age=29&gender=M&e=&zip=11301&...
Host: ad.hi5.com
Referer: http://www.hi5.com/friend/profile/
     displaySameProfile.do?userid=123456789
Cookie: LoginInfo=M_AD_MI_MS|US_0_11301;
     Userid=123456789; Email=jdoe@email.com;
```

Leakage of Age, Gender, Zip and Email Via Request-URI and Cookie to ad.hi5.com, which is a DNS alias for a yieldmanager.com (Yahoo) server.

First-party cookies of hi5.com are being served to this *hidden* third-party server.

Observed direct PII leakage for 2 out of 12 OSNs studied.

Linking PII Leakage

```
GET /pagead/ads?client=ca-primedia-premium_js&...
Host: googleads.g.doubleclick.net
Referer: http://pregnancy.about.com/
Cookie: id=2015bdfb9ec||t=1234359834|et=730|cs=7aepmsks
```

Same cookie as used in accesses to OSNs so doubleclick.net is able to link users to accesses they may like to keep private.

Protection Against PII Leakage

Parties:

1. User

   Could filter out HTTP headers—filtering of cookies is already supported by browser.

   Potential problem with the `Referer` header to leak private information has been known since 1996.

2. Aggregators—filter out PII-related headers. Make cookie semantics more visible.
3. OSNs—could have strong default privacy protection. Easiest is to strip internal user identifier or map user identifier to opaque string on a per-session basis.
4. External applications—could employ one of methods to strip the id or internally remap it.

Leakage via non-OSN sites

Similar manner of leakage could affect users who have accounts and PII on other sites.

Carried out a *preliminary* examination of several popular commercial sites for which we have readily available access.

Included books, newspaper, travel, micropayment, and e-commerce sites.

Identified a news site that leaks user email addresses to at least three separate third-party aggregators.

A travel site embeds a user's first name and default airport in its cookies, which is therefore leaked to any hidden third-party servers.

Did *not* observe leakage of user's login identifier via the `Referer` header, the Cookie, or the Request-URI.

Requires further study.

Summary

Results of study clearly show that the indirect leakage of PII via OSN identifiers to third-party aggregation servers is happening.

OSNs consistently demonstrate leakage of user identifier information to one or more third-parties via Request-URIs, `Referer` headers and cookies.

Also some direct PII leakage.

External applications also leak OSN identifiers.

OSNs in best position to prevent such leakage by eliminating visible OSN identifiers.

Note that aggregators *may* have contractual agreements not to exploit data that they may have access to as a result of actions by users on OSNs.

Future work to study PII leakage to third parties via non-OSN sites.